Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease : a meta-analysis of genome-wide association studies

23 Me Res Team

2019-12

23 Me Res Team , Syst Genomics Parkinson's Dis , Int Parkinson's Dis Genomics , Nalls , M A , Blauwendraat , C & Tienari , P 2019 , ' Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease : a meta-analysis of genome-wide association studies ' , Lancet Neurology , vol. 18 , no. 12 , pp. 1091-1102 . https://doi.org/10.1016/S1474-4422(19)30320-5

http://hdl.handle.net/10138/323872
https://doi.org/10.1016/S1474-4422(19)30320-5

publishedVersion

Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies

Mike A Nalls, Cornelis Blauwendraat*, Costanza L Vallerga*, Karl Heilbron*, Sara Bandres-Ciga*, Diana Chang*, Manuela Tan, Demis A Kia, Alastair J Noyce, Angli Xue, Jose Bras, Emily Young, Rainer von Coelln, Javier Simón-Sánchez, Claudia Schulte, Manu Sharma, Lynne Krohn, Lasse Pihlstrøm, Ari Siitonen, Hirotaka Iwaki, Hampton Leonard, Faraz Faghri, J Raphael Gibbs, Dena G Hernandez, Sonja W Scholz, Juan A Botia, Maria Martinez, Jean-Christophe Corvol, Suzanne Lesage, Joseph Jankovic, Lisa M Shulman, Margaret Sutherland, Pentti Tienari, Kari Majamaa, Mathias Toft, Ole A Andreassen, Tushar Bangale, Alexis Brice, Jian Yang, Ziv Gan-Or, Thomas Gasser, Peter Heutink, Joshua M Shulman, Nicholas W Wood, David A Hinds, John A Hardy, Huw R Morris, Jacob Gratten, Peter M Visscher, Robert R Graham, Andrew B Singleton, on behalf of the 23andMe Research Team†, System Genomics of Parkinson's Disease Consortium†, and the International Parkinson's Disease Genomics Consortium†*

Summary

Background Genome-wide association studies (GWAS) in Parkinson's disease have increased the scope of biological knowledge about the disease over the past decade. We aimed to use the largest aggregate of GWAS data to identify novel risk loci and gain further insight into the causes of Parkinson's disease.

Methods We did a meta-analysis of 17 datasets from Parkinson's disease GWAS available from European ancestry samples to nominate novel loci for disease risk. These datasets incorporated all available data. We then used these data to estimate heritable risk and develop predictive models of this heritability. We also used large gene expression and methylation resources to examine possible functional consequences as well as tissue, cell type, and biological pathway enrichments for the identified risk factors. Additionally, we examined shared genetic risk between Parkinson's disease and other phenotypes of interest via genetic correlations followed by Mendelian randomisation.

Findings Between Oct 1, 2017, and Aug 9, 2018, we analysed 7·8 million single nucleotide polymorphisms in 37 688 cases, 18 618 UK Biobank proxy-cases (ie, individuals who do not have Parkinson's disease but have a first degree relative that does), and 1·4 million controls. We identified 90 independent genome-wide significant risk signals across 78 genomic regions, including 38 novel independent risk signals in 37 loci. These 90 variants explained 16–36% of the heritable risk of Parkinson's disease depending on prevalence. Integrating methylation and expression data within a Mendelian randomisation framework identified putatively associated genes at 70 risk signals underlying GWAS loci for follow-up functional studies. Tissue-specific expression enrichment analyses suggested Parkinson's disease loci were heavily brain-enriched, with specific neuronal cell types being implicated from single cell data. We found significant genetic correlations with brain volumes (false discovery rate-adjusted $p=0 \cdot 0035$ for intracranial volume, $p=0 \cdot 024$ for putamen volume), smoking status ($p=0 \cdot 024$), and educational attainment ($p=0 \cdot 038$). Mendelian randomisation between cognitive performance and Parkinson's disease risk showed a robust association ($p=8 \cdot 00 \times 10^{-7}$).

Interpretation These data provide the most comprehensive survey of genetic risk within Parkinson's disease to date, to the best of our knowledge, by revealing many additional Parkinson's disease risk loci, providing a biological context for these risk factors, and showing that a considerable genetic component of this disease remains unidentified. These associations derived from European ancestry datasets will need to be followed-up with more diverse data.

Funding The National Institute on Aging at the National Institutes of Health (USA), The Michael J Fox Foundation, and The Parkinson's Foundation (see appendix for full list of funding sources).

Lancet Neurol 2019; 18: 1091–102

See Comment page 1069

*Joint first authors

†Full consortia membership is available in the appendix

Laboratory of Neurogenetics, National Institute on Aging (M A Nalls PhD, C Blauwendraat PhD, S Bandres-Ciga PhD, H Iwaki MD, H Leonard MS, F Faghri PhD, J R Gibbs PhD, D G Hernandez PhD, A B Singleton PhD) **and National Institute of Neurological Disorders and Stroke** (S W Scholz MD, M Sutherland PhD)**, National Institutes of Health, Bethesda, MD, USA; Data Tecnica International, Glen Echo, MD, USA** (M A Nalls, H Iwaki, H Leonard)**; Queensland Brain Institute** (A Xue BS, J Yang PhD, P M Visscher PhD)**, Institute for Molecular Bioscience** (C L Vallerga MS, A Xue, J Yang, J Gratten PhD, P M Visscher)**, Mater Research Institute** (J Gratten)**, The University of Queensland, Brisbane, QLD, Australia; 23andMe, Sunnyvale, CA, USA** (K Heilbron PhD, D A Hinds PhD)**; Department of Human Genetics, Genentech, South San Francisco, CA, USA** (D Chang PhD, T Bangale PhD, R R Graham PhD)**; Department of Molecular Neuroscience** (M Tan MS, D A Kia BS, A J Noyce MD, J A Botia PhD, N W Wood MD, J A Hardy)**, Department of Neurodegenerative Diseases** (J Bras)**, Department of Clinical and Movement Neuroscience and UCL Movement Disorders Centre, UCL Queen Square**

Introduction

Parkinson's disease is a neurodegenerative disorder, affecting approximately 1 million individuals in the USA alone.[1] Patients with Parkinson's disease have a combination of progressive motor and non-motor symptoms affecting daily function and quality of life. The prevalence of Parkinson's disease is projected to double in some age groups by 2030, creating a substantial burden on health-care systems.[1]

Early investigations into the role of genetic factors in Parkinson's disease focused on the identification of rare mutations underlying familial disease;[2,3] however, over the past decade there has been a growing appreciation for the contribution of genetics in sporadic disease.[4,5] Genetic

Institute of Neurology, London, UK (H R Morris MD, D A Kia, M Tan, N W Wood); Preventive Neurology Unit, Wolfson Institute of Preventive Medicine, Queen Mary University of London, London, UK (A J Noyce); Center for Neurodegenerative Science, Van Andel Research Institute, Grand Rapids, MI, USA (J Bras PhD); Department of Neurology (E Young, J Jankovic MD, J M Shulman MD), Department of Molecular and Human Genetics (J M Shulman), and Department of Neuroscience (J M Shulman) Baylor College of Medicine, Houston, TX, USA; Department of Neurology, University of Maryland School of Medicine, Baltimore, MD, USA (L M Shulman PhD, R von Coelln MD); Department for Neurodegenerative Diseases, Hertie Institute for Clinical Brain Research (J Simón-Sánchez PhD, C Schulte PhD, T Gasser MD, P Heutink PhD), Centre for Genetic Epidemiology, Institute for Clinical Epidemiology and Applied Biometry (M Sharma PhD), University of Tübingen, Tübingen, Germany; German Center for Neurodegenerative Diseases, Tübingen, Germany (J Simón-Sánchez, C Schulte, T Gasser, P Heutink); Department of Human Genetics (L Krohn MS, Z Gan-Or MD), Montreal Neurological Institute (L Krohn, Z Gan-Or), and Department of Neurology and Neurosurgery (Z Gan-Or), McGill University, Montreal, QC, Canada; Department of Neurology (L Pihlstrøm MD, M Toft MD) and Division of Mental Health and Addiction (O A Andreassen), Oslo University Hospital, Oslo, Norway; Institute of Clinical Medicine, Department of Neurology, University of Oulu, Oulu, Finland (A Siitonen PhD, K Majamaa MD); Department of Neurology and Medical Research Center, Oulu University Hospital, Oulu, Finland (A Siitonen, K Majamaa); The Michael J Fox Foundation, New York, NY, USA (H Iwaki); Department of Computer Science, University of Illinois Urbana-Champaign, Champaign, IL, USA (F Faghri); Department of Neurology,

Research in context

Evidence before this study

Previous studies have used genome-wide association study (GWAS) methods to discover 42 independent risk loci associated with Parkinson's disease. Some of these loci harbouring common risk variants also include rare variants implicated in familial Parkinson's disease risk such as *SNCA*, *LRRK2*, or *GBA*. Earlier studies have attempted to quantify how much heritable risk is captured by common variation that can be easily imputed using commercial genotyping arrays and estimate the amount of risk explained by GWAS. Since 2011, GWAS of Parkinson's disease have integrated expression and methylation datasets to evaluate possible candidate genes for follow-up at Parkinson's disease loci. Many epidemiological and observational studies have attempted to assess risk of Parkinson's disease and various exposures like smoking, caffeine, or occupational hazards, with a mixed track record of success at validating presumed associations.

Added value of this study

This study increased the count of independent common genetic risk factors for Parkinson's disease to 90. We added 38 novel risk variants not previously identified as genome-wide significant. We refined heritability estimates and genetic risk predictions suggesting that common genetic variants account for approximately 22% of Parkinson's disease risk on the liability scale, with a range of 16–36% of that risk being explained by GWAS loci in this study. These updated risk predictions also suggested that polygenic risk scoring can be used to achieve an area under the curve of near 70%, although this prediction uses many more variants than just the 90 independent risk factors identified in this report. Of the 90 risk variants we have characterised here, we have nominated at least one possible candidate gene for follow-up functional studies in 70 of these genomic regions by mining recently available expression and methylation reference datasets on a scale not possible just a few years ago. We have additionally mined single cell RNA sequencing data from mice to identify tissue-specific signatures of enrichment relating to Parkinson's disease genetic risk, showing a major focus on neuronal cell types. We also used the massive amount of publicly available GWAS results to survey genetic correlations between Parkinson's disease and other phenotypes showing significant correlations with smoking, education, and brain morphology. Subsequent analyses using Mendelian randomisation methods showed probable causal links between increased cognitive performance and Parkinson's disease risk on a genetic level.

Implications of all the available evidence

Using updated heritability estimates and risk predictions, we took preliminary steps on a long path to early detection. In future studies, combining genetic and clinicodemographic risk factors could lead to earlier detection and refined diagnostics, which could help improve clinical trials. The generation of copious amounts of public summary statistics created by this effort relating to both the GWAS and subsequent analyses of gene expression and methylation patterns might be of use to investigators planning follow-up functional studies in stem cells or other cellular screens. This information would allow researchers to prioritise targets more efficiently, using our data as additional evidence. We hope our findings might have some downstream clinical impact in the future, such as improved patient stratification for clinical trials and genetically informed drug targets.

studies of sporadic Parkinson's disease have altered the foundational view of disease causes.

We aimed to undertake the largest-to-date GWAS for Parkinson's disease to identify novel mechanistic candidate genes for this disease. We will further assess the function of potential risk genes, estimate Parkinson's disease heritability, and develop a model to predict the proportion of this heritability. Our final goal is to identify potential Parkinson's disease biomarkers and risk factors.

Methods

Study design

The work flow and rationale of our study is shown in figure 1. Three sources of data were used for discovery analyses, these include three previously published GWAS studies,[4,6] 13 new datasets (figure 1), and proxy-case data from the UK Biobank. Previous studies comprised summary statistics published in Nalls and colleagues,[6] GWAS summary statistics from the 23andMe Web-Based Study of Parkinson's Disease by Chang and colleagues,[4] and the publicly available NeuroX dataset from the International Parkinson's Disease Genomics Consortium previously used as a replication sample.[6] These cohorts have been reported in detail, but in brief represent all European ancestry Parkinson's disease case-control GWAS studies available for collaboration.[4,6] We included 13 new case-control sample series for meta-analyses through either publicly available data or collaborations (appendix pp 1–3, 22). All samples from the 13 new datasets underwent similar standardised quality control for inclusion, mirroring that of previous studies. We attempted to generate summary statistics for GWAS meta-analyses as uniformly as possible. This analysis used fixed-effects meta-analyses as implemented in METAL[7] to combine summary statistics across all sources.

Conditional joint analysis

To nominate variants of interest, we used a conditional and joint analysis strategy to algorithmically identify variants that best account for the heritable variation within and across loci.[8] Additional analyses were used to further scrutinise putative associated variants and account for possible differential linkage disequilibrium (LD) signatures, including using the massive single site reference

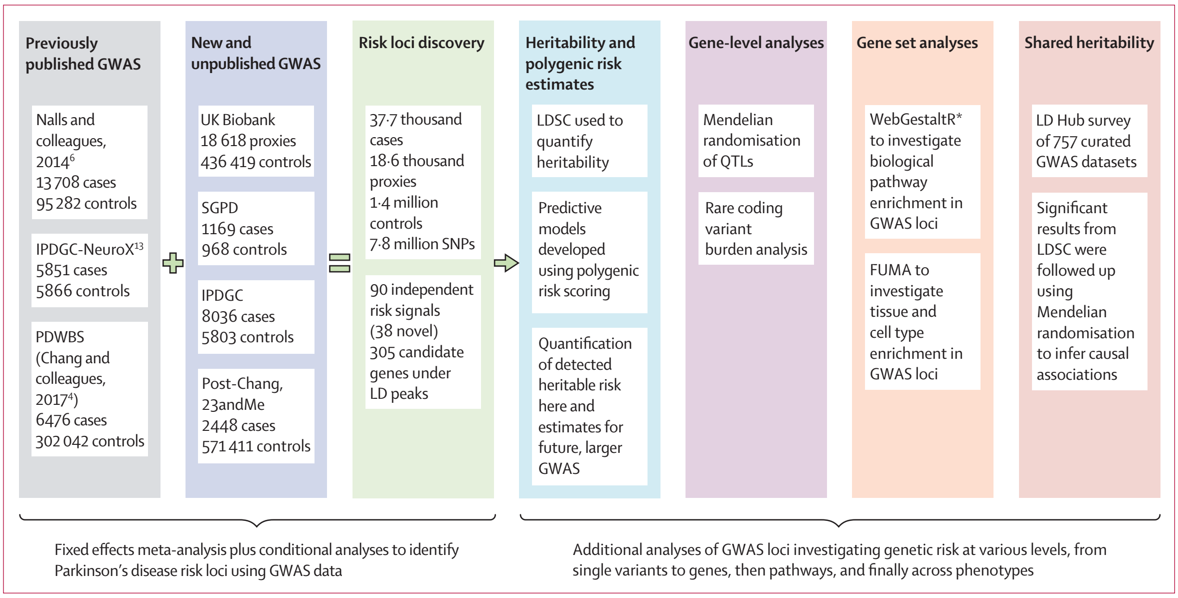

Figure 1: **Workflow and rationale summary**
GWAS=genome-wide association studies. SNPs=single nucleotide polymorphisms. IPDGC=International Parkinson's Disease Genomics Consortium. PDWBS=Parkinson's disease web based study. SGPD= Systems genomics of Parkinson's disease consortium. LD=linkage disequilibrium. LDSC=linkage disequilibrium score regression. QTLs=quantitative trait loci. FUMA=functional mapping and annotation of genetic associations platform. *WEB-based GEne SeT AnaLysis Toolkit.

data from 23andMe in further conditional analyses. If a variant nominated during the conditional and joint analysis strategy phase of analysis was greater than 1 Mb from any of the genome-wide significant loci nominated in Chang and colleagues,[4] we considered this to be a novel risk variant. We defined nominated risk variants as from a single locus if they were within 250 kb of each other. We instituted two filters after fixed-effects and conditional and joint analyses, excluding variants that had a random-effects p value across all datasets more than $4 \cdot 67 \times 10^{-4}$ and a conditional analysis p value more than $4 \cdot 67 \times 10^{-4}$ using participant level 23andMe genotype data. Please see appendix (pp 4–5, 22) summarising all variants nominated.

Heritability estimates and extant genetic risk

We used the R package PRSice2 for risk profiling,[9] which calculates polygenic risk score (PRS) profiling in the standard weighted allele dose manner.[4,6,10–12] In addition, PRSice incorporates permutation testing, in which case and control labels are swapped in the withheld samples to generate an empirical p value. This workflow identifies the best p value thresholds for variant inclusion while doing LD pruning. In many cases this best p value threshold for PRS construction does not meet what is commonly regarded as genome-wide significance.

A two-stage design was also used, training on the largest single array study (NeuroX-dbGaP[13]) and then tested on the second largest study (Harvard Biomarker Study) using the same array. These two targeted array studies were chosen for three reasons: precedent in the previous publications in which the NeuroX-dbGaP dataset was used in PRS; direct genotyping of larger effect rare variants in *GBA* and *LRRK2*; and participant level genotypes for these datasets are publicly available.

To calculate heritability in clinically defined Parkinson's disease datasets, we used LD score regression employing the LD references for Europeans provided with the software.[14] This workflow was also repeated on a per cohort level (appendix pp 10–11).

Functional causal inferences via quantitative trait loci

We used Mendelian randomisation to test whether changes in DNA methylation or RNA expression of genes physically proximal to significant Parkinson's disease risk loci were causally related to Parkinson's disease risk. To nominate genes of interest for Mendelian randomisation analyses, we took our putative 90 loci in the large LD reference used for the conditional and joint phase of analysis and identified SNPs in LD with our SNPs at an r^2 0·5 within 1 Mb (appendix pp 9–10). Mendelian randomisation was used by integrating discovery phase summary statistics with quantitative trait loci (QTL) association summary statistics across well curated methylation and expression datasets. We used the curated versions of Qi and colleagues brain methylation and expression summary statistics (multistudy and multi-tissue meta-analysis), with a specific focus on substantia nigra data (GTEx). We also made use of the blood expression data from Võsa and colleagues from 2018 (eQTLGen).[15–19] For all QTL analyses, we used the multi-SNP summary-based Mendelian randomisation method as a framework to do Mendelian randomisation. All Mendelian randomisation effect estimates are reported on the scale of an SD increase in the exposure variable relating

Johns Hopkins University Medical Center, Baltimore, MD, USA (S W Scholz); **Departamento de Ingeniería de la Información y las Comunicaciones, Universidad de Murcia, Spain** (J A Botia); **Institut national de la santé et de la recherche médicale Unité mixte de recherche 1220, Toulouse, France** (M Martinez PhD); **Paul Sabatier University, Toulouse, France** (M Martinez PhD); **Institut national de la santé et de la recherche médicale U1127, CNRS UMR 7225, Paris, France** (J-C Corvol PhD, S Lesage PhD, A Brice MD); **Sorbonne Université centre national de la recherche médicale, unité mixte de recherche 1127, Paris, France** (J-C Corvol PhD, S Lesage PhD, A Brice MD); **Assistance Publique Hôpitaux de Paris, Paris, France** (J-C Corvol PhD, S Lesage PhD, A Brice MD); **Institut du Cerveau et de la Moelle épinière, Paris, France** (J-C Corvol, S Lesage, A Brice) **Clinical Neurosciences, Neurology, University of Helsinki, Helsinki, Finland** (P Tienari MD); **Helsinki University Hospital, Helsinki, Finland** (P Tienari); **Institute of Clinical Medicine** (M Toft) **and Jebsen Centre for Psychosis Research** (O A Andreassen MD), **University of Oslo, Oslo, Norway; and Jan and Dan Duncan Neurological Research Institute, Texas Children's Hospital, Houston, TX, USA** (J M Shulman)

Correspondence to:
Dr Mike A Nalls, Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Bethesda, MD 20892, USA
mike@datatecnica.com

See **Online** for appendix

For **WebGestaltR** see http://www.webgestalt.org/

For the **conditional and joint analysis strategy** see http://cnsgenomics.com/software/gcta/

For the **Harvard Biomarker Study** see https://neurodiscovery.harvard.edu/biomarkers-discovery

For **GTEx** see http://cnsgenomics.com/software/smr/#DataResource

For **eQTLGen** see http://www.eqtlgen.org to a similar change in Parkinson's disease risk. Simply, these Mendelian randomisation analyses compare the local polygenic risk of an exposure (methylation or expression) to similar polygenic risk in an outcome (Parkinson's disease). This method infers causal associations under the assumption that there is no intermediate confounder associated with both parameters and that the association is not simply due to LD.

For the **Oxford Brain Imaging Genetics server** see http://big.stats.ox.ac.uk/

To further investigate expression enrichment across cell types in Parkinson's disease, we integrated GWAS summary statistics with expression and network data from the Functional Mapping and Annotation of Genome-Wide Association Studies (FUMA) webserver.[20]

For the **FUMA webserver version 1.3.1** see https://fuma.ctglab.nl/

Rare coding variant burden tests

A uniformly quality controlled and imputed dataset from the International Parkinson's Disease Genomics Consortium was used to do burden tests for all rarer coding variants successfully imputed in a mean of 85% of the sample series (17 188 cases and 22 875 controls). These analyses include all variants at a hard call threshold of imputation quality more than 0·8. After annotation with ANNOVAR, we had 37 503 exonic coding variants (non-synonymous, stop, or splicing) at minor allele frequency less than 5% and a subset of 29 016 at minor allele frequency less than 1%.[21] For inclusion in this phase, a gene had to contain at least two coding variants. After assembling this subset of 113 testable genes, we used the optimised sequence kernel association test to generate summary statistics at maximum minor allele frequencies of 1% and 5%.[22]

For **ANNOVAR** see http://annovar.openbioinformatics.org/en/latest/

LD score regression and causal inference

To investigate correlations of Parkinson's disease genetics with that of multiple traits and diseases, we used bivariate LD score regression.[14] These analyses were done with data from the 757 GWAS available via LD Hub and biomarker GWAS summary statistics[23–25] on c-reactive protein and cytokine measures; LD Hub was accessed on June 20, 2018, (version 1.2.0).[23–25] The p values from the bivariate LD score regression were adjusted for false discovery rate to account for multiple testing. Traits showing significant genetic correlations with Parkinson's disease were analysed with Mendelian randomisation methods. We excluded the UK Biobank data when a nominated trait was from summary statistics derived from the UK Biobank or if the UK Biobank was included as part of a meta-analysis.

For **LD Hub** see http://ldsc.broadinstitute.org/ldhub/

When complete GWAS summary statistics were available for traits of interest (relating to smoking and education), we used the more powerful bidirectional generalised summary-data-based Mendelian randomisation. We analysed GWAS summary statistics for smoking initiation (453 693 records from a self-report survey with 208 988 regular smokers and 244 705 never regular smokers) and current smoking within the UK Biobank, current smoking contrasted 47 419 current smokers versus 244 705 never regular smokers. The same analysis was done incorporating recent GWAS data regarding educational attainment (N=766 345) from self-report in the UK and cognitive performance (N=257 828) as measured by the *g* composite score.[26] These data were analysed using methods to mirror that of the UK Biobank Parkinson's disease GWAS dataset. Combined left and right putamen volume from a T2 weighted MRI GWAS was available from Oxford Brain Imaging Genetics server (accessed Dec 28, 2018).[27] All Mendelian randomisation analyses included GWAS on the scale of ten thousand samples and overcame the considerable power demands of the methods. For additional quality control, method details, and ancillary results, see the appendix.

Role of the funding source

The funder of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. The corresponding author had full access to all of the data and the final responsibility to submit for publication.

Results

Our study took place between Oct 1, 2017, and Aug 9, 2018. To maximise our power for locus discovery we used a single stage design, meta-analysing all available GWAS summary statistics. Supporting this design, we found strong genetic correlations using Parkinson's disease cases ascertained by clinicians compared with 23andMe self-reported cases (genetic correlation from LD score regression [*r*G] 0·85, SE 0·06) and UK Biobank proxy cases (*r*G 0·84, SE 0·134).

We identified 90 independent genome-wide significant association signals through our analyses of 37 688 cases, 18 618 UK Biobank proxy-cases, and 1 417 791 controls at 7 784 415 SNPs (figure 2, table 1, appendix pp 1–6). Of these, 38 signals are newly identified and more than 1 Mb from loci described previously (appendix p 4).[4]

We detected ten loci containing more than one independent risk signal (22 risk SNPs in total across these loci), of which nine had been identified by previous GWAS, including multi-signal loci in the vicinity of *GBA, NUCKS1* and *RAB29, GAK* and *TMEM175, SNCA,* and *LRRK2*. The novel multi-signal locus comprised independent risk variants rs2269906 (*UBTF* and *GRN*) and rs850738 (*FAM171A2*). Detailed summary statistics on all nominated loci can be found in the appendix (pp 4–6), including variants filtered out during additional quality control.

To quantify how much of the genetic liability we have explained and what direction to take with future Parkinson's disease GWAS, we generated updated heritability estimates and PRS. Using LD score regression on a meta-analysis of all 11 clinically ascertained datasets from our GWAS, we estimated the liability-scale heritability of Parkinson's disease as 0·22 (95% CI 0·18–0·26), only slightly lower than a previous estimate derived using genome-wide complex trait analysis (0·27, 0·17–0·38).[14,28,29]

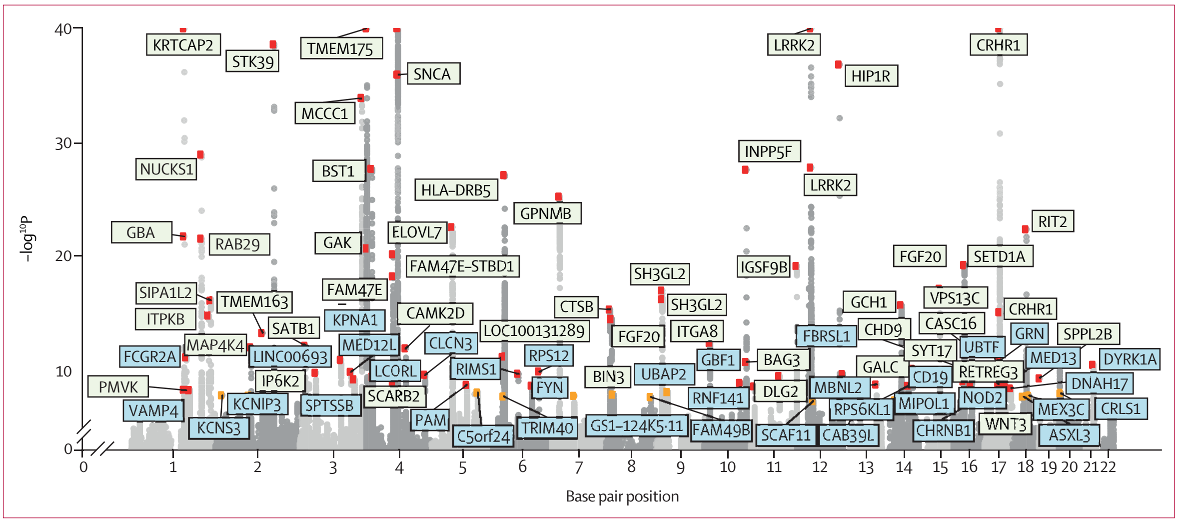

***Figure 2:* Manhattan plot for significant variants**

The nearest gene to each of the 90 significant variants are labelled in green for previously identified loci and in blue for novel loci. $-\log_{10}$ p values were capped at 40. Variant points are colour-coded red and orange, with orange representing significant variants at $p=5\times10^{-8}$ and 5×10^{-9} and red representing significant variants at $p<5\times10^{-9}$. The X axis represents the base pair position of variants from smallest to largest per chromosome (1–22), only autosomes were included in this analysis.

LD score regression is known to be more conservative than genome-wide complex trait analysis; however, our LD score regression heritability estimate does fall within the 95% CI of this estimate.

To establish the proportion of SNP-based heritability explained by our Parkinson's disease GWAS results with PRS, we used a two-stage design, with variant selection and training in the NeuroX-dbGaP dataset[13] (5851 cases and 5866 controls) and then validation in the Harvard Biomarker Study (527 cases and 472 controls). Using equations from Wray and colleagues[30] and our current heritability estimates, the 88 variant PRS explained a minimum 16% of the genetic liability of Parkinson's disease assuming a global prevalence of 0·5%.[28,30] The 1805 variant PRS explained 26% of Parkinson's disease heritability. In a high-risk population with a prevalence of 2%, the 1805 variant PRS explained a maximum 36% of Parkinson's disease heritable risk (appendix pp 6–7).[28,30]

We then attempted to quantify strata of risk in our more inclusive PRS. Compared with individuals with PRS values in the lowest quartile, the Parkinson's disease odds ratio for individuals with PRS values in the highest quartile was 3·74 (95% CI 3·35–4·18) in the NeuroX-dbGaP cohort and 6·25 (4·26–9·28) in the Harvard Biomarker Study cohort (table 2, figure 3, appendix p 16).

Variants with p values in the range of 5×10^{-8} to $1{\cdot}35\times10^{-3}$ (used in the 1805 variant PRS) were rarer and had smaller effect estimates than variants reaching genome-wide significance. These sub-significant variants had a median minor allele frequency of 21·3% and a median effect estimate (absolute value of the log odds ratio of the SNP parameter from regression) of 0·047. Genome-wide significant risk variants were more common with a median minor allele frequency of 25·1%, and had a median effect estimate of 0·081. Here we assume that the lower minor allele frequencies and smaller effect size estimates are typical and representative of variants contributing to our more inclusive PRS and represent future GWAS hits. We did power calculations to forecast the number of additional Parkinson's disease cases needed to achieve genome-wide significance at 80% power for a variant with a minor allele frequency of 21·3% and an effect estimate of 0·047.[31] Assuming that future data is well harmonised with current data and that disease prevalence is 0·5%, we estimated that we would need around 99 000 cases, around 2·3 times more than the present study for these to reach genome-wide significance. These variants already contribute towards the current increases in area under the curve (AUC) when considering the 1805 variant PRS outperforms the 88 variant PRS. Expanding future studies to this size will invariably identify new loci and improve the AUC for a genetic predictor in Parkinson's disease (maximum potential AUC estimated at 85% using the equations from Wray and colleagues[30]).

There were 305 genes within the 78 GWAS loci. We sought to identify the probable causal gene or genes in each locus using large QTL datasets and summary-data-based Mendelian randomisation (table 3, appendix pp 8–9).[17] This method allows for functional inferences between two datasets to be made in an analogous framework to a randomised controlled trial, treating the genotype as the randomising factor.

Of the 305 genes under LD peaks around our risk variants of interest, 237 were possibly associated with at least one QTL in public reference datasets and were therefore testable via summary-based Mendelian randomisation (appendix pp 8–9). The expression or methylation of 151 (64%) of these 237 genes was significantly associated with a possible causal change in Parkinson's disease risk.

	Chromosome	Base pair position	Nearest gene	Effect allele	Other allele	Effect allele frequency	OR (95% CI)	Regression coefficient (β)	Standard error of β	p value, fixed-effects	p value, conditional joint analysis approach	p value, conditional	p value, random-effects	I^2, %
rs6658353	1	161469054	*FCGR2A*	C	G	0·501	1·07 (1·05–1·09)	0·065	0·009	$6·10\times10^{-12}$	$4·69\times10^{-12}$	$1·38\times10^{-5}$	$3·71\times10^{-5}$	40·2%
rs11578699	1	171719769	*VAMP4*	T	C	0·195	0·93 (0·91–0·95)	–0·070	0·012	4·47×10–9	4·45×10–9	2·63×10–3	1·09×10–7	5·1%
rs76116224	2	18147848	*KCNS3*	A	T	0·904	1·12 (1·08–1·16)	0·110	0·019	1·27×10–8	1·27×10–8	3·75×10–7	1·27×10–8	0
rs2042477	2	96000943	*KCNIP3*	A	T	0·242	0·94 (0·92–0·96)	–0·066	0·012	$1·38\times10^{-8}$	$1·48\times10^{-8}$	$3·49\times10^{-5}$	$1·38\times10^{-8}$	0
rs6808178	3	28705690	*LINC00693*	T	C	0·379	1·07 (1·05–1·09)	0·066	0·010	$8·09\times10^{-12}$	$7·18\times10^{-12}$	$8·84\times10^{-5}$	$8·09\times10^{-12}$	0
rs55961674	3	122196892	*KPNA1*	T	C	0·172	1·09 (1·06–1·12)	0·086	0·013	$9·98\times10^{-12}$	$8·30\times10^{-12}$	$2·80\times10^{-6}$	$9·98\times10^{-12}$	0
rs11707416	3	151108965	*MED12L*	A	T	0·367	0·94 (0·92–0·96)	–0·063	0·010	$1·13\times10^{-10}$	$1·02\times10^{-10}$	$2·66\times10^{-4}$	$1·77\times10^{-7}$	10·9%
rs1450522	3	161077630	*SPTSSB*	A	G	0·674	0·94 (0·92–0·96)	–0·062	0·010	$5·01\times10^{-10}$	$4·90\times10^{-10}$	$3·51\times10^{-4}$	$2·27\times10^{-5}$	24·6%
rs34025766	4	17968811	*LCORL*	A	T	0·159	0·92 (0·90–0·94)	–0·084	0·013	$2·87\times10^{-10}$	$2·82\times10^{-10}$	$7·43\times10^{-6}$	$2·87\times10^{-10}$	0
rs62333164	4	170583157	*CLCN3*	A	G	0·326	0·94 (0·92–0·96)	–0·064	0·010	$2·00\times10^{-10}$	$1·77\times10^{-10}$	$5·10\times10^{-5}$	$2·17\times10^{-5}$	21·3%
rs26431	5	102365794	*PAM*	C	G	0·703	1·06 (1·04–1·09)	0·062	0·010	$1·57\times10^{-9}$	$1·65\times10^{-9}$	$6·00\times10^{-3}$	$2·36\times10^{-7}$	7·9%
rs11950533	5	134199105	*C5orf24*	A	C	0·102	0·91 (0·88–0·94)	–0·092	0·016	$7·16\times10^{-9}$	$6·73\times10^{-9}$	$5·08\times10^{-4}$	$2·68\times10^{-8}$	1·9%
rs9261484	6	30108683	*TRIM40*	T	C	0·245	0·94 (0·92–0·96)	–0·064	0·011	$1·62\times10^{-8}$	$1·43\times10^{-8}$	$1·26\times10^{-6}$	$1·62\times10^{-8}$	0
rs12528068	6	72487762	*RIMS1*	T	C	0·284	1·07 (1·05–1·09)	0·066	0·010	$1·63\times10^{-10}$	$1·79\times10^{-10}$	$9·80\times10^{-6}$	$1·63\times10^{-10}$	0
rs997368	6	112243291	*FYN*	A	G	0·805	1·07 (1·05–1·10)	0·071	0·012	$1·84\times10^{-9}$	$1·97\times10^{-9}$	$2·61\times10^{-5}$	$1·84\times10^{-9}$	0
rs75859381	6	133210361	*RPS12*	T	C	0·967	0·80 (0·75–0·86)	–0·221	0·034	$1·04\times10^{-10}$	$9·67\times10^{-11}$	$1·09\times10^{-6}$	$1·04\times10^{-10}$	0
rs76949143	7	66009851	*GS1-124K5·11*	A	T	0·051	0·87 (0·82–0·91)	–0·143	0·025	$1·43\times10^{-8}$	$1·51\times10^{-8}$	$5·47\times10^{-9}$	$2·04\times10^{-6}$	12·3%
rs2086641	8	130901909	*FAM49B*	T	C	0·723	0·94 (0·92–0·96)	–0·061	0·011	$1·81\times10^{-8}$	$1·57\times10^{-8}$	$6·07\times10^{-6}$	$1·81\times10^{-8}$	0
rs6476434	9	34046391	*UBAP2*	T	C	0·734	0·94 (0·92–0·96)	–0·062	0·011	$6·58\times10^{-9}$	$6·56\times10^{-9}$	$2·74\times10^{-4}$	$6·58\times10^{-9}$	0
rs10748818	10	104015279	*GBF1*	A	G	0·851	0·92 (0·90–0·95)	–0·079	0·013	$1·05\times10^{-9}$	$1·23\times10^{-9}$	$7·47\times10^{-6}$	$1·05\times10^{-9}$	0
rs7938782	11	10558777	*RNF141*	A	G	0·878	1·09 (1·06–1·12)	0·087	0·015	$2·12\times10^{-9}$	$1·97\times10^{-9}$	$2·17\times10^{-7}$	$2·12\times10^{-9}$	0
rs7134559	12	46419086	*SCAF11*	T	C	0·404	0·95 (0·93–0·97)	–0·054	0·010	$3·96\times10^{-8}$	$3·80\times10^{-8}$	$1·69\times10^{-2}$	$1·84\times10^{-5}$	25·2%
rs11610045	12	133063768	*FBRSL1*	A	G	0·490	1·06 (1·04–1·08)	0·060	0·009	$1·77\times10^{-10}$	$1·62\times10^{-10}$	$3·57\times10^{-5}$	$8·79\times10^{-7}$	19·5%
rs9568188	13	49927732	*CAB39L*	T	C	0·740	1·06 (1·04–1·09)	0·062	0·011	$1·15\times10^{-8}$	$1·11\times10^{-8}$	$4·29\times10^{-6}$	$2·46\times10^{-4}$	21·4%
rs4771268	13	97865021	*MBNL2*	T	C	0·230	1·07 (1·05–1·09)	0·068	0·011	$1·45\times10^{-9}$	$1·67\times10^{-9}$	$1·41\times10^{-4}$	$1·45\times10^{-9}$	0
rs12147950	14	37989270	*MIPOL1*	T	C	0·438	0·95 (0·93–0·97)	–0·053	0·010	$3·54\times10^{-8}$	$3·58\times10^{-8}$	$1·06\times10^{-3}$	$3·54\times10^{-8}$	0
rs3742785	14	75373034	*RPS6KL1*	A	C	0·787	1·07 (1·05–1·10)	0·071	0·012	$1·92\times10^{-9}$	$2·08\times10^{-9}$	$2·22\times10^{-6}$	$8·18\times10^{-6}$	24·8%
rs2904880	16	28944396	*CD19*	C	G	0·309	0·94 (0·92–0·96)	–0·065	0·011	$7·87\times10^{-10}$	$8·68\times10^{-10}$	$1·39\times10^{-5}$	$7·87\times10^{-10}$	0
rs6500328	16	50736656	*NOD2*	A	G	0·599	1·06 (1·04–1·08)	0·059	0·010	$1·82\times10^{-9}$	$1·53\times10^{-9}$	$1·43\times10^{-3}$	$1·82\times10^{-9}$	0
rs12600861	17	7355621	*CHRNB1*	A	C	0·648	0·95 (0·93–0·96)	–0·057	0·010	$1·01\times10^{-8}$	$1·15\times10^{-8}$	$5·10\times10^{-3}$	$1·01\times10^{-8}$	0
rs2269906	17	42294337	*UBTF*	A	C	0·653	1·07 (1·04–1·09)	0·063	0·010	$6·24\times10^{-10}$	$8·63\times10^{-9}$	$1·17\times10^{-5}$	$6·24\times10^{-10}$	0
rs850738	17	42434630	*FAM171A2*	A	G	0·606	0·93 (0·91–0·95)	–0·071	0·011	$1·29\times10^{-11}$	$3·55\times10^{-10}$	$4·18\times10^{-4}$	$2·17\times10^{-7}$	17·0%
rs61169879	17	59917366	*BRIP1*	T	C	0·164	1·09 (1·06–1·11)	0·082	0·013	$9·28\times10^{-10}$	$9·40\times10^{-10}$	$9·07\times10^{-7}$	$6·21\times10^{-6}$	16·4%
rs666463	17	76425480	*DNAH17*	A	T	0·833	1·08 (1·05–1·11)	0·076	0·013	$3·20\times10^{-9}$	$2·90\times10^{-9}$	$1·62\times10^{-5}$	$4·17\times10^{-4}$	41·0%
rs1941685	18	31304318	*ASXL3*	T	G	0·498	1·05 (1·04–1·07)	0·053	0·009	$1·69\times10^{-8}$	$1·61\times10^{-8}$	$1·64\times10^{-8}$	$1·69\times10^{-8}$	0
rs8087969	18	48683589	*MEX3C*	T	G	0·550	0·94 (0·93–0·96)	–0·058	0·010	$1·41\times10^{-8}$	$1·46\times10^{-8}$	$1·09\times10^{-4}$	$1·41\times10^{-8}$	0
rs77351827	20	6006041	*CRLS1*	T	C	0·128	1·08 (1·05–1·11)	0·080	0·014	$8·87\times10^{-9}$	$7·94\times10^{-9}$	$1·84\times10^{-5}$	$4·38\times10^{-7}$	11·2%
rs2248244	21	38852361	*DYRK1A*	A	G	0·283	1·07 (1·05–1·10)	0·071	0·011	$2·74\times10^{-11}$	$2·51\times10^{-11}$	$6·31\times10^{-5}$	$8·78\times10^{-6}$	34·3%

Summary statistics for 38 novel genome-wide significant Parkinson's disease variants using data from all available genome-wide association studies.

Table 1: **Novel loci associated with Parkinson's disease**

	Max p value threshold	Pseudo r^2 from PRS*	Regression coefficient (β)	SE of β	p value	OR, highest quartile PRS	95% CI, highest quartile PRS	SNPs (N)	Samples (N)	Area under the curve	95% CI (DeLong)	Sensitivity	Specificity	Positive predictive value	Negative predictive value	Balanced accuracy
Training dataset: IPDGC, Neurox	$1·35\times10^{-3}$	0·029	0·553	0·022	$8·99\times10^{-135}$	3·74	3·35–4·18	1809	11 243	0·640	0·630–0·650	0·569	0·632	0·591	0·611	0·601
Validation dataset: HBS	$4·00\times10^{-2}$	0·054	0·709	0·072	$8·28\times10^{-23}$	6·25	4·26–9·28	1805	999	0·692	0·660–0·725	0·628	0·686	0·691	0·623	0·657

Estimates of performance for predictive models including single study estimates, estimates from meta-analyses across studies, as well as a two stage design. Here the best p value threshold column denotes the filtering value for SNP inclusion to achieve the maximal pseudo (Nagelkerke's) r^2. The OR column is the exponent of the regression coefficient (β) from logistic regression of the PRS on case status, with the SE representing the precision of these estimates. These same metrics are derived across array types and datasets using random-effects meta-analyses. The area under the curve is included as the most common metric for predictive model performance. PRS=polygenic risk score. OR=odds ratio. IPDGC=International Parkinson's Disease Genomics Consortium. HBS=Harvard Biomarker Study. *r^2 approximation adjusted for an estimated prevalence of 0·5%, equivalent to roughly half of the unadjusted r^2 estimates for the PRS. All calculations and reported statistics include only the PRS and no other parameters after adjusting for principal components 1–5, age, and sex at variant selection in the NeuroX-dbGaP dataset[13]. SNP=single nucleotide polymorphism.

Table 2: **Summary of genetic predictive model performance**

Of the 90 Parkinson's disease GWAS risk variants, 70 were in loci containing at least one of these putatively causal genes after multiple test correction (table 3). For 53 (76%) of these 70 Parkinson's disease GWAS hits, the gene nearest to the most significant SNP was a putatively causal gene (appendix pp 4–6). Most loci tested contained multiple putatively causal genes. The nearest putatively causal gene to the rs850738 and *FAM171A2* GWAS risk signal is *GRN*, a gene known to be associated with frontotemporal dementia.[32] Mutations in *GRN* have also been shown to be connected with another lysosomal storage disorder, neuronal ceroid lipofuscinosis.[33]

As an orthogonal approach for nominating genes under GWAS peaks, we did rare coding variant burden analyses. We did kernel-based burden tests on the 113 genes of the 305 under our GWAS peaks that contained two or more rare coding variants (minor allele frequency <5% or <1%). After Bonferroni correction for 113 genes, we identified seven significant genes: *LRRK2*, *GBA*, *CATSPER3* (rs11950533 and *C5orf24* locus), *LAMB2* (rs12497850 and *IP6K2* locus), *LOC442028* (rs2042477 and *KCNIP3* locus), *NFKB2* (rs10748818 and *GBF1* locus), and *SCARB2* (rs6825004 locus). These results suggest that some of the risk associated with these loci might be due to rare coding variants or that these are pleomorphic risk loci. The *LRRK2* and *NFKB2* associations at minor allele frequency less than 1% remained significant after correcting for the approximate 20 000 genes in the human genome ($p=2·15\times10^{-12}$ for *LRRK2* and $p=4·02\times10^{-7}$ *NFKB2*, appendix pp 8–9).

We tested whether genes of interest were enriched in 10 651 biological pathways (from gene ontology annotations) using FUMA.[20,34] We found ten significantly enriched pathways (false discovery rate-adjusted $p<0·05$, appendix pp 9–10), including four related to vacuolar function and three related to known drug targets (calcium transporters, ikeda_mir1_targets_dn and ikeda_mir30_targets_up; kinase signalling, kim_pten_targets_dn[35]). At least three candidate genes within novel loci are involved in lysosomal storage disorders (*GUSB*, *GRN*, and *NEU1*), a pathway of keen interest in Parkinson's disease.[36] Our GWAS results also include candidate genes *VAMP4* and *NOD2* from the endocytic pathway.[37]

To establish the tissues and cell types most relevant to the causes of Parkinson's disease using FUMA,[20,34] we tested whether the genes highlighted by our Parkinson's disease GWAS were enriched for expression in 53 tissues. We found 13 significant tissues, all of which were brain-derived (appendix p 17), in contrast to what has been seen in Alzheimer's disease, which shows a strong bias towards blood, spleen, lung, and microglial enrichments.[38] To further disentangle the enrichment in brain tissues, we tested whether our Parkinson's disease GWAS genes were enriched for expression in 88 brain cell types using single cell RNA sequencing reference data from mouse brains using DropViz.[39] After false discovery rate correction we found seven significant brain cell types, all of which were

For **DropViz** see http://dropviz.org

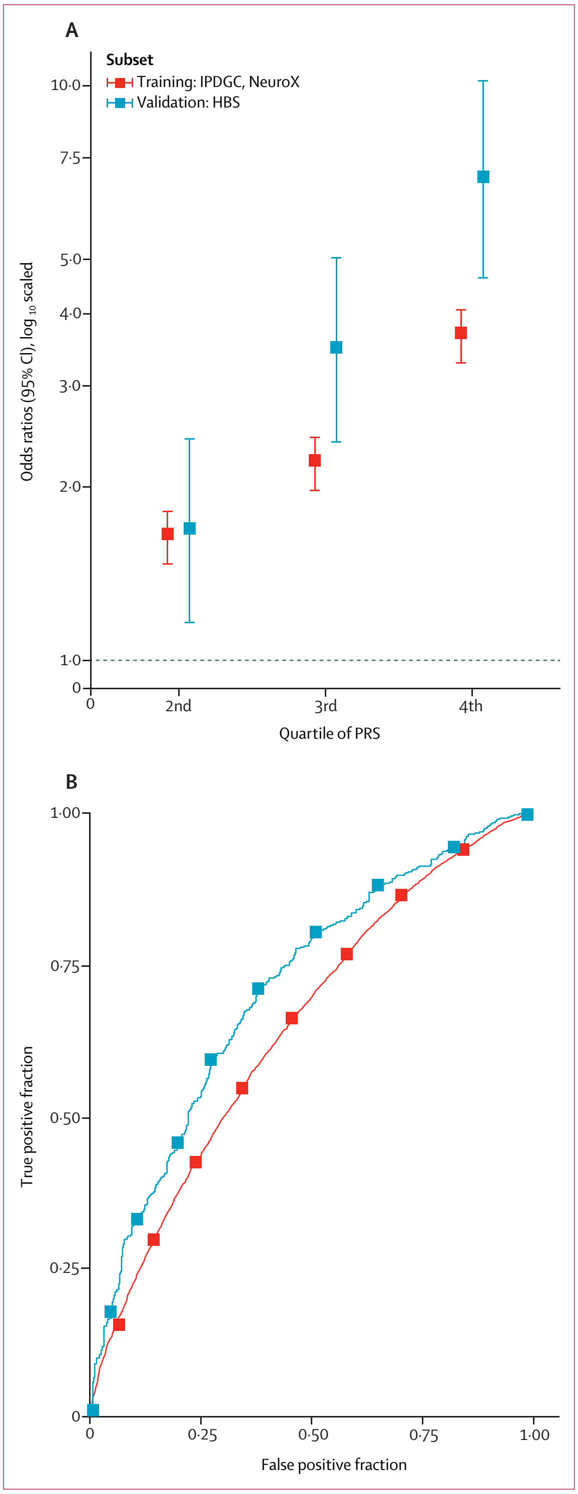

Figure 3: **Predictive model**
The odds ratio of developing Parkinson's disease for each quartile of PRS compared with the lowest quartile of genetic risk (A). PRS receiver-operator curves for the more inclusive 1805 variant PRS in the validation dataset and in the corresponding training dataset that was used for PRS thresholding and single nucleotide polymorphism selection (B). PRS=polygenic risk score. IPDGC=International Parkinson's Disease Genomics Consortium. HBS=Harvard Biomarker Study.

neuronal (appendix p 17). The strongest enrichment was for neurons in the substantia nigra at $p=1\cdot0\times10^{-6}$, with additional significant results at $p<5\cdot0\times10^{-4}$ for the globus pallidus, thalamus, posterior cortex, frontal cortex, hippocampus, and entopeduncular nucleus.

Next, we used cross-trait genetic correlation and Mendelian randomisation to identify possible Parkinson's disease biomarkers and risk factors by comparing with 757 other GWAS datasets curated by LD Hub.[25] We found four significant genetic correlations (false discovery rate-adjusted p<0·05, appendix pp 10–11) including positive correlations with intracranial volume (p=0·0035) and putamen volume (p=0·024),[40] and negative correlations with current tobacco use (p=0·024) and academic qualifications (p=0·038; eg, National Vocational Qualifications, Higher National Diploma, Higher National Certificate, or equivalent).[41] The negative association with an individual's academic qualifications suggests that individuals without a college education might be at less risk of Parkinson's disease. The correlation between Parkinson's disease and smoking status might not be independent from the correlation between Parkinson's disease and education as smoking status and years of education were significantly correlated.[42]

We used Mendelian randomisation to assess whether there was evidence of a causal relationship between Parkinson's disease and five phenotypes related to academic qualifications, smoking, and brain volumes described above (appendix pp 19–21). Cognitive performance had a large, significant causal effect on Parkinson's disease risk (Mendelian randomisation effect 0·213, SE 0·041; Bonferroni-adjusted $p=8\cdot00\times10^{-7}$), whereas Parkinson's disease risk did not have a significant causal effect on cognitive performance (Bonferroni-adjusted p=0·125). Educational attainment also had a significant causal effect on Parkinson's disease risk (Mendelian randomisation effect 0·162, SE 0·040, Bonferroni-adjusted $p=2\cdot06\times10^{-4}$), and Parkinson's disease risk also had a weak but significant causal effect on educational attainment (Mendelian randomisation effect 0·007, SE 0·002, Bonferroni-adjusted $p=7\cdot45\times10^{-3}$). There was no significant causal relationship between Parkinson's disease and current smoking status in forward analysis (Mendelian randomisation effect –0·069, SE 0·031; Bonferroni-adjusted p=0·125) or reverse analysis (Mendelian randomisation effect 0·004, SE 0·010, Bonferroni-adjusted p=1·00). Smoking initiation (the act of ever starting smoking) did not have a causal effect on Parkinson's disease risk (Mendelian randomisation effect –0·063, SE 0·034, Bonferroni-adjusted p=0·32), whereas Parkinson's disease had a small, but significantly positive causal effect on smoking initiation (Mendelian randomisation effect 0·027, SE 0·006, Bonferroni-adjusted $p=1\cdot62\times10^{-5}$). Intracranial volume could not be tested because the GWAS data (available from Oxford Brain Imaging Genetics server) did not contain any genome-wide significant risk variants. No significant causal relationship was observed between Parkinson's disease and

	Probe	Chromosome	Probe, base pair	Top SNP, base pair	Top SNP	SNPs (N)	Effect	SE	p value	Bonferroni adjusted p value
VAMP4[19]	ENSG00000117533	1	171 690 343	171 717 417	rs10913587	98	–0·272	0·05	$5·67 \times 10^{-7}$	$1·19 \times 10^{-4}$
KCNIP3[16]	ENSG00000115041	2	96 007 438	95 989 766	rs3772034	14	–0·161	0·04	$1·12 \times 10^{-5}$	$1·15 \times 10^{-3}$
MAP4K4[19]	ENSG00000071054	2	102 410 880	102 338 377	rs6733355	3	1·119	0·24	$2·32 \times 10^{-6}$	$4·87 \times 10^{-4}$
TMEM163[16]	ENSG00000152128	2	135 344 950	135 248 544	rs598668	28	0·074	0·02	$3·55 \times 10^{-7}$	$3·65 \times 10^{-5}$
KPNA1[19]	ENSG00000114030	3	122 187 294	122 201 610	rs73190142	110	0·310	0·05	$1·56 \times 10^{-6}$	$3·28 \times 10^{-4}$
GAK[16]	ENSG00000178950	4	884 612	906 131	rs11248057	1	0·508	0·10	$7·47 \times 10^{-7}$	$7·69 \times 10^{-5}$
CAMK2D[19]	ENSG00000145349	4	114 527 635	114 730 260	rs115671064	146	–0·006	0·05	$5·74 \times 10^{-6}$	$1·21 \times 10^{-3}$
PAM[19]	ENSG00000145730	5	102 228 247	102 118 633	rs2432162	679	–0·031	0·01	$2·08 \times 10^{-6}$	$4·36 \times 10^{-4}$
LOC100131289[16]	cg21339923	6	27 636 378	27 636 378	rs78149975	2	–0·094	0·02	$1·53 \times 10^{-6}$	$3·06 \times 10^{-4}$
TRIM40[16]	cg01641092	6	30 094 300	30 094 315	rs9261443	8	0·072	0·01	$6·15 \times 10^{-6}$	$1·23 \times 10^{-3}$
HLA-DRB5[16]	cg26036029	6	32 552 443	32 570 311	rs34039593	8	–0·153	0·02	$7·53 \times 10^{-10}$	$1·51 \times 10^{-7}$
GPNMB[16]	ENSG00000136235	7	23 295 156	23 294 668	rs858274	74	0·090	0·01	$2·73 \times 10^{-21}$	$2·81 \times 10^{-19}$
CTSB[16]	ENSG00000164733	8	11 713 495	11 699 279	rs4631423	33	–0·150	0·04	$4·37 \times 10^{-9}$	$4·50 \times 10^{-7}$
BIN3[16]	ENSG00000147439	8	22 502 296	22 456 517	rs71513892	32	0·046	0·01	$1·43 \times 10^{-6}$	$1·48 \times 10^{-4}$
SH3GL2[16]	ENSG00000107295	9	17 688 103	17 684 784	rs10756899	15	0·252	0·05	$5·83 \times 10^{-8}$	$6·00 \times 10^{-6}$
ITGA8[16]	ENSG00000077943	10	15 659 036	15 548 925	rs7910668	6	–0·201	0·05	$6·13 \times 10^{-5}$	$6·32 \times 10^{-3}$
RNF141[19]	ENSG00000110315	11	10 548 001	10 553 355	rs4910153	120	–0·054	0·05	$6·25 \times 10^{-7}$	$1·31 \times 10^{-4}$
IGSF9B[16]	cg25790212	11	133 800 774	133 800 477	rs11223626	1	–0·172	0·04	$3·24 \times 10^{-6}$	$6·48 \times 10^{-4}$
FBRSL1[16]	cg03621470	12	133 137 479	133 138 334	rs10781619	16	–0·057	0·01	$6·35 \times 10^{-5}$	$1·27 \times 10^{-2}$
CAB39L[16]	ENSG00000102547	13	49 950 524	49 918 175	rs35214871	30	0·097	0·02	$3·51 \times 10^{-8}$	$3·62 \times 10^{-6}$
GCH1[16]	ENSG00000131979	14	55 339 148	55 348 837	rs3825611	6	0·113	0·03	$2·76 \times 10^{-4}$	$2·85 \times 10^{-2}$
SYT17[16]	ENSG00000103528	16	19 229 472	19 273 554	rs727747	4	0·177	0·05	$1·54 \times 10^{-4}$	$1·58 \times 10^{-2}$
SETD1A[19]	ENSG00000099381	16	30 982 526	30 950 352	rs7206511	34	–0·710	0·09	$2·75 \times 10^{-13}$	$5·77 \times 10^{-11}$
CHRNB1[16]	ENSG00000170175	17	7 354 703	7 373 595	rs60488855	18	0·115	0·03	$1·67 \times 10^{-5}$	$1·72 \times 10^{-3}$
UBTF[19]	ENSG00000108312	17	42 290 697	42 297 631	rs113844752	34	–0·466	0·09	$5·68 \times 10^{-6}$	$1·19 \times 10^{-3}$
MAPT[16]	ENSG00000186868	17	44 038 724	44 862 347	rs199502	6	0·265	0·03	$7·13 \times 10^{-24}$	$7·35 \times 10^{-22}$
WNT3 (GTEx v7)	ENSG00000108379·5	17	44 875 148	44 908 263	rs9904865	2	–0·082	0·02	$4·01 \times 10^{-6}$	$4·81 \times 10^{-5}$
DNAH17[16]	cg09006072	17	76 425 972	76 427 732	rs589582	3	0·100	0·02	$2·44 \times 10^{-5}$	$4·88 \times 10^{-3}$
MEX3C[19]	ENSG00000176624	18	48 722 797	48 731 131	rs12458916	40	–0·291	0·05	$5·28 \times 10^{-5}$	$1·11 \times 10^{-2}$

Multi-SNP eQTL Mendelian randomisation results focusing only on the most significant association per nearest gene to the Parkinson's disease risk variant of interest after Bonferroni correction. If a locus was significantly associated with both brain and blood QTLs after multiple test correction, we opted to show the most signficant brain tissue derived association here after filtering for possible polygenicity (HEIDI p>0·01). Effects with significant HEIDI p values might indicate a possible effect complicated by linkage disequilibrium and are less likely to be a true causal association. All tested QTL summary statistics can be found in the appendix pp 8–9. Effect estimates represent the change in Parkinson's disease odds ratio per one SD increase in gene expression or methylation. SNP=single nucleotide polymorphism. QTL=quantitative trait locus.

Table 3: **Summary of significant functional inferences from QTL associations via Mendelian randomisation for nominated genes of interest**

putamen volume ($p>0\cdot05$ in both the forward and reverse directions).

Discussion

This meta-analysis of GWAS marks a crucial step forward in our understanding of the genetic architecture of Parkinson's disease and provides a genetic reference set for the broader research community. We identified 90 independent common genetic risk factors for Parkinson's disease, nearly doubling the number of known Parkinson's disease risk variants. We re-evaluated the cumulative contribution of genetic risk variants, both of genome-wide significance and not yet discovered, to refine our estimates of heritable Parkinson's disease risk. We also nominated probable genes at each locus for further follow-up using QTL analyses and rare variant burden analyses. Our work has highlighted the pathways, tissues, and cell types involved in Parkinson's disease causes. Finally, we identified intracranial and putaminal volume as potential future Parkinson's disease biomarkers, and cognitive performance as a Parkinson's disease risk factor. Altogether, the data presented here has substantially expanded the resources available for future investigations into potential Parkinson's disease interventions.

We were able to explain 16–36% of Parkinson's disease heritability, the range being directly related to varying prevalence estimates (0·5–2·0%). Power estimates suggest that expansions of case numbers to 99 000 cases will continue to reveal additional insights into Parkinson's disease genetics. Although these risk variants will have small effects or be quite rare, they will help to further expand our knowledge of the genes and pathways that drive Parkinson's disease risk.

Population-wide screening for individuals who are likely to develop Parkinson's disease is currently not feasible using our 1805 variant PRS alone. There would be roughly 14 false positives per true positive assuming a prevalence of 0·5%. Although large-scale genome sequencing and non-linear machine learning methods will probably improve these predictive models, we have previously shown that we will need to incorporate other data sources (eg, smell tests, family history, age, sex) to generate algorithms that have more value in population-wide screening.[43]

Evaluating these results in the larger context of pathway, tissue, and cellular functionality revealed that genes near Parkinson's disease risk variants showed enrichment for expression in the brain, contrasting with previous findings in Alzheimer's disease. Strikingly, we showed that the expression enrichment of genes at Parkinson's disease loci occurred exclusively in neuronal cell types. We also found that Parkinson's disease genes were enriched in chemical signalling pathways and pathways involving the response to a stressor. We believe that this contrast, in which the pathway enrichment analyses suggest at least some immune component to Parkinson's disease and the expression enrichment analysis does not suggest any significant immune related tissue component should be viewed with caution. In particular, the marginal p values of most immune-related pathways in our analyses after multiple test correction reinforce this caution. These observations could be informative for disease modelling efforts, highlighting the importance of disease modelling in neurons and possibly incorporating a cellular stress component. This information can help inform and focus stem-cell derived therapeutic development efforts that are underway.[44]

We found four phenotypes that were genetically correlated with Parkinson's disease. Putamen and intracranial volumes might prove to be valuable in future Parkinson's disease biomarker studies. Our bidirectional generalised summary-data-based Mendelian randomisation results suggest a complex causative connection between smoking initiation and Parkinson's disease that will require further follow-up. One of the implications of this work is that Parkinson's disease trials of nicotine or other smoking-related compounds might be less likely to succeed due to the marginal strength of associations shown in this report affecting study power. The strong causal effect of cognitive performance on Parkinson's disease is supported by observational studies.[45]

Although this study marks major progress in assessing genetic risk factors for Parkinson's disease, much remains to be established. No defined external validation dataset was used, which could be seen as a limitation. However, due to the size of the dataset, it is considered infeasible to build a sufficiantly large replication series. Also, external replication of the novel associations we present will be difficult simply because of the sample sizes needed. Simulations have suggested that, without replication, variants with p values between 5×10^{-8} and 5×10^{-9} should be interpreted with greater caution.[46,47] We found 16 risk variants in this range, including two known variants near *WNT3* (proximal to the *MAPT* locus) and *BIN3*. To a degree, the fact that we filtered our variants with a secondary random-effects meta-analysis could make our 90 Parkinson's disease GWAS hits somewhat more robust because of the conservative nature of random-effects.

This study focused on Parkinson's disease risk in individuals of European ancestry. Adding datasets from non-European populations would be helpful to further improve our granularity in association testing and ability to fine-map loci through integration of more variable LD signatures while also evaluating population-specific associations. Also, risk predictions might not generalise across populations in some cases and ancestry specific PRS should be investigated. Additionally, large ancestry-specific Parkinson's disease LD reference panels, such as those for patients who are Ashkenazi Jews, will help us further unravel the genetic risk of loci such as *GBA* and *LRRK2*. This clarification might be particularly crucial at these loci where LD patterns could be variable within European populations, accentuating the possible influence of LD reference series on conditional analyses in some cases.[48] Finally, our work used state-of-the-art QTL datasets to nominate candidate genes, but many QTL associations are hampered by both small sample size and low *cis*-SNP density. Larger QTL studies and Parkinson's disease-specific network data from large scale cellular screens would allow us to build a more robust functional inference framework.

As the field moves forward there are some crucial next steps that should be prioritised. First, allowing researchers to share participant-level data in a secure environment would facilitate inclusiveness and uniformity in analyses while maintaining the confidentiality of study participants. Our work suggests that GWAS of increasing size will continue to provide useful biological insights into Parkinson's disease. In addition to studies of the genetics of Parkinson's disease risk, studies of disease onset, progression, and subtype will be important and will require large series of well characterised patients.[49] We also believe that work across diverse populations is important, not only to be able to best serve these populations but also to aid in fine mapping of loci. Notably, the use of genome sequencing technologies could further improve discovery by capturing rare variants and structural variants, but with the caveat that very large sample sizes will be required. Although much is left to do, we believe that our study represents a substantial step forward and that the results and data will serve as a foundational resource for the community to pursue this next phase of Parkinson's disease research.

Contributors

MAN, CB, CLV, KH, SB-C, DC, MT, DK, LR, JS-S, LK, LP, and ABS were responsible for study-level analysis. MAN, CB, SB-C, AJN, AX, JY, JG, PMV, and ABS were responsible for additional analysis and data management. MAN, CB, CLV, KH, SB-C, LP, MS, KM, MT, AB, JY, ZG-O, TG, PH, JMS, NW, DAH, JH, HRM, JG, PMV, RRG, and ABS were responsible for design and funding. MAN, CB, CLV, KH, SB-C, DC, MT, DAK, AJN, AX, JB, EY, RvC, JS-S, CS, MS, LK, LP, AS, HI, HL, FF, JRG, DGH, SWS, JAB, MM, OAA, J-CC, SL, JJ, LMS, MS, PT, KM, MT, AB, JY, ZG-O, TG, PH, JMS, NW, DAH, JH, HRM, JG, PMV, RRG, and ABS were responsible for critical review and writing of the manuscript.

Declaration of interests

MAN reports that this work was done under a consulting contract with National Institutes of Health, he also consults for Lysosomal Therapeutics Inc, Neuron23, and Illumina. KH reports other support from 23andMe, during the conduct of the study outside the submitted work. DC reports other support from Genentech outside the submitted work. MT reports grants from Parkinson's UK during the conduct of the study. AJN reports grants from Parkinson's UK and Virginia Kieley benefaction; grants and non-financial support from GE Healthcare; and personal fees from Profile, Bial, and Britannia, outside the submitted work. RvC reports grants from American Brain Foundation and Michael and Eugenia Brin Foundation, during the conduct of the study.
LP reports grants from Southeastern Regional Health Authority, Norway, during the conduct of the study. AS reports grants from Sigrid Juselius Foundation, during the conduct of the study. SWS is a scientific advisory council member for the Lewy Body Dementia Association. J-CC reports grants from French Ministry of Health, during the conduct of the study; grants from Sanofi, personal fees from Ever Pharma, Denali, Biogen, Air Liquide, BrainEver, and Theranexus, outside the submitted work. LMS reports grants from National Institutes of Health, outside the submitted work. PT has a patent c9orf72 in the diagnosis and treatment of neurodegenerative disease pending. MT reports grants from Research Council of Norway, Regional Health Authority South-Eastern Norway, and Michael J Fox Foundation, and personal fees from Roche, outside the submitted work. OAA reports grants from Research Council of Norway, and KG Jebsen Stiftelsen during the conduct of the study; and personal fees from Lundbeck, outside the submitted work; OAA also has a patent (PCT/US2014/011014) pending. TB reports other from Genentech, outside the submitted work. AB reports grants from France Parkinson and Fondation pour la Recherche sur le Cerveau, grants from Agence nationale de recherche (ANR) EPIG, grants from ANR joint program in neurodegenerative disease, grants from Roger de Spoelberch Foundation, France Alzheimer, Ecole des neurosciences Paris, Institut de France, CHU de Nimes, European Rare Disease Network, ANR EPIG, and APHP, outside the submitted work. ZG-O reports personal fees from Lysosomal Therapeutics, Idorsia, Inception Sciences, Denali, and Prevail Therapeutics, outside the submitted work. TG reports grants from The Michael J Fox Foundation for Parkinson´s Research, personal fees from UCB Pharma, other from Joint Programming for Neurodegenerative Diseases program, funded by the European Commission, personal fees from Novartis, Teva, and MedUpdate, outside the submitted work; and has a patent issued (EP1802749 [A2] KASPP [LRRK2] gene), for the detection and treatment of neurodegenerative disorders issued PH is a consultant for NEURON23. JMS reports grants from Burroughs Wellcome Fund during the conduct of the study, grants from National Institutes of Health, and personal fees from Helis Medical Foundation, outside the submitted work. DAH reports other from 23andMe, outside the submitted work. HRM reports grants from NIHR Biomedical Research Centre, Parkinson's UK, Medical Research Council , and The Cure Parkinson's Trust, during the conduct of the study; grants from PSP Association and CBD Solutions, personal fees from Teva, Boehringer Ingelheim, GlaxoSmithKline, grants from Drake Foundation, personal fees from Biogen, UCB, and Biohaven, outside the submitted work. RRG reports other from Genentech, during the conduct of the study and other from Genentech, outside the submitted work. All other authors declare no competing interests.

Data sharing

GWAS summary statistics for the post-Chang 23andMe dataset and 23andMe summary statistics included in the studies of Chang and colleagues[4] and Nalls and colleagues[6] will be made available through 23andMe to qualified researchers under an agreement with 23andMe that protects the privacy of the 23andMe participants. Interested investigators should visit http://research.23andme.com/dataset-access to submit a request. An immediately accessible version of the summary statistics is available online, excluding Nalls and colleagues[6], 23andMe post-Chang and colleagues[4] and Web-Based Study of Parkinson's Disease but including all analysed SNPs. After approval from 23andMe, the full summary statistics including all analysed SNPs and samples in this GWAS meta-analysis will be accessible to approved researchers. Underlying participant level International Parkinson's Disease Genomics Consortium data are available to potential collaborators, please contact ipdgc.contact@gmail.com.

For **23andMe publication dataset access and for more information** see https://research.23andme.com/collaborate/#publication

For **summary statistics** see https://bit.ly/2ofzGrk

Acknowledgments

For details of individual funding see appendix p 16.

References

1 Dorsey ER, Constantinescu R, Thompson JP, et al. Projected number of people with Parkinson disease in the most populous nations, 2005 through 2030. *Neurology* 2007; **68:** 384–86.

2 Polymeropoulos MH, Higgins JJ, Golbe LI, et al. Mapping of a gene for Parkinson's disease to chromosome 4q21-q23. *Science* 1996; **274:** 1197–99.

3 Singleton AB, Farrer M, Johnson J, et al. α-Synuclein locus triplication causes Parkinson's disease. *Science* 2003; **302:** 841.

4 Chang D, Nalls MA, Hallgrímsdóttir IB, et al. A meta-analysis of genome-wide association studies identifies 17 new Parkinson's disease risk loci. *Nat Genet* 2017; **49:** 1511–16.

5 Fung H-C, Scholz S, Matarin M, et al. Genome-wide genotyping in Parkinson's disease and neurologically normal controls: first stage analysis and public release of data. *Lancet Neurol* 2006; **5:** 911–16.

6 Nalls MA, Pankratz N, Lill CM, et al. Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease. *Nat Genet* 2014; **46:** 989–93.

7 Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* 2010; **26:** 2190–91.

8 Yang J, Ferreira T, Morris AP, et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat Genet* 2012; **44:** 369–75, S1–3.

9 Euesden J, Lewis CM, O'Reilly PF. PRSice: polygenic risk score software. *Bioinformatics* 2014; **31:** 1466–68.

10 International Parkinson Disease Genomics Consortium, et al. Imputation of sequence variants for identification of genetic risks for Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet* 2011; **377:** 641–49.

11 International Parkinson's Disease Genomics Consortium (IPDGC), Wellcome Trust Case Control Consortium 2 (WTCCC2). A two-stage meta-analysis identifies several new loci for Parkinson's disease. *PLoS Genet* 2011; **7:** e1002142.

12 Nalls MA, Escott-Price V, Williams NM, et al. Genetic risk and age in Parkinson's disease: continuum not stratum. *Mov Disord* 2015; **30:** 850–54.

13 Nalls MA, Bras J, Hernandez DG, et al .NeuroX, a fast and efficient genotyping platform for investigation of neurodegenerative diseases. Neurobiol Aging 2015; 36: e7–12.

14 Bulik-Sullivan BK, Loh P-R, Finucane HK, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* 2015; **47:** 291–95.

15 GTEx Consortium. The genotype-tissue expression (GTEx) project. *Nat Genet* 2013; **45:** 580–85.

16 Qi T, Wu Y, Zeng J, et al. Identifying gene targets for brain-related traits using transcriptomic and methylomic data from blood. *Nat Commun* 2018; **9:** 2282.

17 Zhu Z, Zhang F, Hu H, et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat Genet* 2016; **48:** 481–87.

18 Wu Y, Zeng J, Zhang F, et al. Integrative analysis of omics summary data reveals putative mechanisms underlying complex traits. *Nat Commun* 2018; **9:** 918.

19 Võsa U, Claringbould A, Westra H-J, et al. Unraveling the polygenic architecture of complex traits using blood eQTL meta-analysis. *bioRxiv* 2018; published online Oct 19. DOI:10.1101/447367 (preprint).

20 Watanabe K, Taskesen E, van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. *Nat Commun* 2017; **8:** 1826.

21 Yang H, Wang K. Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. *Nat Protoc* 2015; **10:** 1556–66.

22 Lee S, Emond MJ, Bamshad MJ, et al. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am J Hum Genet* 2012; **91:** 224–37.

23 LD Hub. http://ldsc.broadinstitute.org/ldhub/ (accessed Jun 20, 2018).

24 Prins BP, Abbasi A, Wong A, et al. Investigating the causal relationship of c-reactive protein with 32 complex somatic and psychiatric outcomes: a large-scale cross-consortium Mendelian randomization study. *PLoS Med* 2016; **13:** e1001976.

25 Ahola-Olli AV, Würtz P, Havulinna AS, et al. Genome-wide association study identifies 27 loci influencing concentrations of circulating cytokines and growth factors. *Am J Hum Genet* 2017; **100:** 40–50.

26 Lee JJ, Wedow R, Okbay A, et al. Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1·1 million individuals. *Nat Genet* 2018; published online Jul 23. DOI:10.1038/s41588-018-0147-3.

27 Elliott LT, Sharp K, Alfaro-Almagro F, et al. Genome-wide association studies of brain imaging phenotypes in UK Biobank. *Nature* 2018; **562:** 210–16.

28 Keller MF, Saad M, Bras J, et al. Using genome-wide complex trait analysis to quantify "missing heritability" in Parkinson's disease. *Hum Mol Genet* 2012; **21:** 4996–5009.

29 Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* 2011; **88:** 76–82.

30 Wray NR, Yang J, Goddard ME, Visscher PM. The genetic interpretation of area under the ROC curve in genomic profiling. *PLoS Genet* 2010; 6: e1000864.

31 Skol AD, Scott LJ, Abecasis GR, Boehnke M. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat Genet* 2006; **38:** 209–13.

32 Cruts M, Gijselinck I, van der Zee J, et al. Null mutations in progranulin cause ubiquitin-positive frontotemporal dementia linked to chromosome 17q21. *Nature* 2006; **442:** 920–24.

33 Smith KR E al. Strikingly different clinicopathological phenotypes determined by progranulin-mutation dosage. *Am J Hum Genet* 2012; **90:** 1102–07.

34 Wang J, Vasaikar S, Shi Z, Greer M, Zhang B. WebGestalt 2017: a more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit. *Nucleic Acids Res* 2017; **45:** W130–37.

35 Gene set enrichment analysis. 2019. http://software.broadinstitute.org/gsea/index.jsp (accessed June 20, 2018).

36 Robak LA, Jansen IE, van Rooij J, et al. Excessive burden of lysosomal storage disorder gene variants in Parkinson's disease. *Brain* 2017; **140:** 3191–203.

37 Bandres-Ciga S, Saez-Atienzar S, Bonet-Ponce L, et al. The endocytic membrane trafficking pathway plays a major role in the risk of Parkinson's disease. *Mov Disord* 2019; **34:** 460–68.

38 Jansen IE, Savage JE, Watanabe K, et al. Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nat Genet* 2019; **51:** 404–13.

39 Saunders A, Macosko EZ, Wysoker A, et al. Molecular diversity and specializations among the cells of the adult mouse brain. *Cell* 2018; **174:** 1015–30

40 Hibar DP, Stein JL, Renteria ME, et al. Common genetic variants influence human subcortical brain structures. *Nature* 2015; **520:** 224–29.

41 Neale lab. Rapid GWAS of thousands of phenotypes for 337,000 samples in the UK Biobank. 2017. http://www.nealelab.is/blog/2017/7/19/rapid-gwas-of-thousands-of-phenotypes-for-337000-samples-in-the-uk-biobank (accessed Jun 24, 2018).

42 Bulik-Sullivan B, ReproGen Consortium, Finucane HK, Anttila V, Gusev A, Day FR. An atlas of genetic correlations across human diseases and traits. *Nat Genet* 2015; **47:** 1236–41.

43 Nalls MA, McLean CY, Rick J, et al. Diagnosis of Parkinson's disease on the basis of clinical and genetic classification: a population-based modelling study. *Lancet Neurol* 2015; **14:** 1002–09.

44 Parkinson's progression Markers Inititative. 2019. http://www.ppmi-info.org/access-data-specimens/ (accessed June 20, 2018).

45 Valdés EG, Andel R, Sieurin J, et al. Occupational complexity and risk of Parkinson's disease. *PLoS One* 2014; **9:** e106676.

46 Wu Y, Zheng Z, Visscher PM, Yang J. Quantifying the mapping precision of genome-wide association studies using whole-genome sequencing data. *Genome Biol* 2017; **18:** 86.

47 Pulit SL, de With SAJ, de Bakker PIW. Resetting the bar: statistical significance in whole-genome sequencing-based association studies of global populations. *Genet Epidemiol* 2017; **41:** 145–51.

48 Rivas MA, Avila BE, Koskela J, et al. Insights into the genetic epidemiology of Crohn's and rare diseases in the Ashkenazi Jewish population. *PLoS Genet* 2018; **14:** e1007329.

49 Iwaki H, Blauwendraat C, Leonard HL, et al. Genome-wide association study of Parkinson's disease progression biomarkers in 12 longitudinal patients' cohorts. *bioRxiv* 2019; published online March 27. DOI:10.1101/585836 (preprint).